A modeling strategy for developing gene identifier CDEs

A modeling strategy is presented that provides an approach for integrating all the available collections of biomedical information that will be represented as annotation services on the **ca**ncer **B**iomedical **I**nformatics **G**rid (caBIG). This strategy follows the recommendations made in the CDE white paper prepared by Rakesh Nagarajan of the Genome Annotation Special Interest Group within the Integrative Cancer Research (ICR) Workspace. The whitepaper is available at http://cabig.nci.nih.gov/workspaces/ICR/Meetings/SIGs/gene_annotation/Gene%20CDE%20Focus%20Group/20041007_GeneCDE_whitepaper

Recommendations

The white paper recommends using the **ca**ncer **D**ata **S**tandards **R**epository (caDSR) to store Unified Modeling Language (UML) models of interrelated objects comprising of ISO/IEC 11179 compliant **C**ommon **D**ata **E**lements (CDEs). It also recommends leveraging the controlled vocabularies provided by the **E**nterprise **V**ocabulary **S**ervices (EVS) to define the data element concepts and value domains that combine to form the CDEs. The following steps are outlined for the ICR project developers:

- Each project must describe its objects using a UML model whose classes and attributes are described by terms in the EVS..
- This model must be imported into the caDSR using the UML loader, whereby data elements are created.
- The same data elements representing gene identifiers should be used across projects.

Using the CDEs across different projects facilitates syntactic and semantic integration of data. However, the presence of a common identifier for genes and gene products across all publicly available databases (externally, for e.g. Entrez Gene ID vs Ensembl Gene Identifier, or even internally, for e.g. LocusLinkID vs UniGene ID) poses further complications. The white paper, therefore, makes the following recommendations:

- A list of required CDEs representing a gene or its mRNA or protein product should be gathered by examining the data models of the current ICR projects and by scrutinizing potential future use cases.
- This list of CDEs should either be reused from the existing CDEs in the caDSR or should be created by the ICR community
- Each ICR project's data model should utilize **at least** one or more of these defined CDEs.
- Because the goal is not to be restrictive, if data models in the future cannot reasonably accommodate one of the existing CDEs, additional CDEs may be added to this dynamic list. The Architecture Workspace, Vocabulary/CDE Workspace, and genome annotation subject matter experts (potentially the Genome Annotation SIG members) should oversee the addition of such CDEs.

In keeping up with these recommendations, the following UML modeling approach is proposed that provides a flexible solution to deal with the problem of multiple identifiers for the same objects that is in sync with the long term proposal of providing a mapping service for the gene and gene product identifiers.

The Identifier Class

This approach leaves the onus of developing gene/protein/mRNA/other classes on the individual developers. They are only constrained by having to reference one or more of a set of attributes of a general identifier class. This identifier class contains an extensible attribute list of approved gene identifiers. The CDEs developed from this identifier class can be reused across multiple projects/domains and provide a means of interlinking the various data models. This method does not impose the development of all possible biological classes upon the Gene CDE Focus Group as this list could prove restrictive for certain non-mainstream (albeit important) projects. The idea is to provide flexibility to the individual developers to come up with relevant classes for their projects, interlink them with gene identifiers from this identifier class where appropriate (thus promoting reuse of CDEs) and thus make them recognizable across the grid. If the approved identifier list is not sufficient, new attributes can be added to the identifier class. If the individual developers need to define specific attributes for their classes, they can develop appropriate CDEs and add them to the caDSR repository (if not present). If the biological classes use a common identifier between them, a join can be made on this common element. Also, when a mapping service is developed in the future, objects of the identifier class could be instantiated and filled with mappings across various databases.

An illustrative example:

Identifier Class:
Attributes:
- LocusLink ID
- UniGene ID
- Ensembl Gene ID
- RefSeq mRNA Accession [Set of one or more]
- RefSeq protein Accession [Set of one or more]
- UniProt ID

The following is an example of the type of Data Element that would result from this approach:

Data Element= Object Class + Property + Value Domain
Object class: Identifier
Property: LocusLink ID
Value domain: String

WashU Gene class:
Attributes:
- Reference to Identifier Class Object

- Gene Symbol (e.g. EGR1 [All caps for human])
- Gene Name
- Gene Ontology Associations [Set]

Georgetown Protein class:
Attributes:
- Reference to Identifier Class Object
- Protein Symbol (e.g. *EGR1* [all caps AND italicized since it is a protein])
- Posttranslational modifications

If these classes employ a common identifier (for e.g. LocusLink ID), they can be joined on this CDE.

Caveats

The transformation of the UML models developed using this approach presents a problem as the caDSR currently does not have a means of imposing a constraint that the biological classes should use at least one of a set of identifiers. However, this feature needs to be incorporated into the UML Loader at some point of time.

Conclusions

The modeling approach described above facilitates:

- Flexibility
- Extensibility
- Reuse of CDEs

While the caveat mentioned above places some hindrances initially in the loading of UML models developed using this approach into the caDSR, the impetus for this issue to be resolved sooner rather than later is provided by the advantages this approach offers in terms of satisfying the primary goal of interlinking various ICR projects within the grid and the promise it offers in terms of providing a viable mapping service in the future.